

10/652,653

日 本 国 特 許 庁
JAPAN PATENT OFFICE

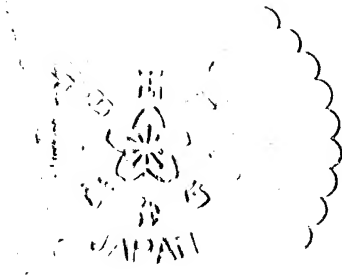
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 1 月 1 4 日
Date of Application:

出 願 番 号 特 願 2 0 0 3 - 0 0 5 2 3 4
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 0 0 5 2 3 4]

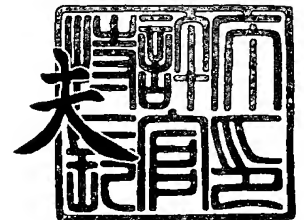
出 願 人 株式会社日立製作所
Applicant(s):



2 0 0 3 年 8 月 2 6 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 0 6 9 6 2 7

【書類名】 特許願

【整理番号】 H02018281A

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 13/00

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 田中 淳

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 橋本 顕義

【発明者】

 【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 R A I D システム事業部内

 【氏名】 平児 典夫

【特許出願人】

 【識別番号】 000005108

 【氏名又は名称】 株式会社 日立製作所

【代理人】

 【識別番号】 100075096

 【弁理士】

 【氏名又は名称】 作田 康夫

 【電話番号】 03-3212-1111

【手数料の表示】

 【予納台帳番号】 013088

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1



【物件名】

図面 1

【物件名】

要約書 1

【プルーフの要否】

要

【書類名】 明細書

【発明の名称】 SAN/NAS統合型ストレージ装置

【特許請求の範囲】

【請求項 1】

データを記憶する複数のディスクと、
ホストコンピュータからのI/O命令を受信し、上記I/O命令に従って上記ディスクを制御する複数の制御装置と、
前記複数の制御手段を結合する相互結合網とを有し、
上記制御装置の1つは、ホストコンピュータとのインタフェースとして、ブロックインタフェースを持ち、
上記制御装置の別の1つはホストコンピュータとのインタフェースとして、ファイルインタフェースを持つことを特徴とするSAN/NAS統合型ストレージ装置。

【請求項 2】

上記ブロックインタフェースを持つ制御装置は、
上記ホストコンピュータと通信をするチャネルアダプタと、
上記ディスクを制御するディスクアダプタとを有する請求項 1 記載のSAN/NAS統合型ストレージ装置。

【請求項 3】

上記ファイルインタフェースを持つ制御装置は、
上記ホストコンピュータとの通信をするファイルサーバ部と、
上記ホストコンピュータから受信したファイルレベルのコマンド、データをブロックインタフェースに変換するファイルシステムと、
上記相互結合網とのインターフェイスを持つチャネルアダプタ部と、
上記ディスクを制御するディスクアダプタとを有する請求項 1 記載のSAN/NAS統合型ストレージ装置。

【請求項 4】

上記相互結合網に接続され、上記複数の制御装置で共用される制御メモリを有する請求項 1 記載のSAN/NAS統合型ストレージ装置。

【請求項 5】

上記ファイルインタフェースを持つ制御装置は、
上記ホストコンピュータとの通信をするファイルサーバ部と、
上記ホストコンピュータから受信したファイルレベルのコマンド、データをブロックインタフェースに変換するファイルシステムと、
上記相互結合網と上記のインターフェイスを持つチャネルアダプタ部を、
単一のボード上に実装することを特徴とする請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 6】

上記ファイルインタフェースを持つ制御装置は、
上記ホストコンピュータとの通信をするファイルサーバ部と、
上記ホストコンピュータから受信したファイルレベルのコマンド、データをブロックインタフェースに変換するファイルシステムと、
上記ファイルシステムから発行されたブロックレベルのコマンド、データを受信するチャネルアダプタ部と、
上記ディスクを制御するディスクアダプタとを有する請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 7】

上記ファイルインタフェースを持つ制御装置は、
上記ホストコンピュータとの通信をするファイルサーバ部と、
ブロックレベルの入出力命令、入出力データを受信するチャネルアダプタ部間に通信パスを複数持ち、コマンド、データを別々のパスにて送受信することを特徴とする請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 8】

上記ファイルインタフェースを持つ制御装置は、
上記ホストコンピュータとの通信をするファイルサーバ部と、ブロックレベルの入出力命令、入出力データを受信するチャネルアダプタ部間にある物理的な通信パスを、仮想的に複数の独立したパスが存在するように制御し、コマンド、データを別々のパスにて送受信することを特徴とする請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 9】

上記ファイルインタフェースを持つ制御装置は、

上記ホストコンピュータとの通信をするファイルサーバ部と、ブロックレベルの入出力命令、入出力データを受信するチャネルアダプタ部間にある物理的な通信パスを、仮想的に複数の独立したパスが存在するように制御し、コマンド内部にファイルシステムに固有な情報を挿入し、コマンドとデータを別々のパスにて送受信することを特徴とする請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 10】

上記ファイルインタフェースを持つ制御装置は、

上記ホストコンピュータとの通信をするファイルサーバ部と、ブロックレベルの入出力命令、入出力データを受信するチャネルアダプタ部間に障害情報用の通信パスを持ち、コマンド、データの通信パスとは独立に障害情報を制御メモリへ送信することを特徴とする請求項 1 記載の SAN/NAS 統合型ストレージ装置。

【請求項 11】

データを記憶するディスクと、

第 1 のホストコンピュータに接続される第 1 のネットワークに直結される第 1 のアダプタと、

上記ディスクに直結される第 2 のアダプタと、

上記第 1 のアダプタ、第 2 のアダプタに直結してこれらを結合する第 2 のネットワークとを有し、

上記第 1 のアダプタは、上記第 1 のネットワークを介してホストコンピュータから受信した第 1 のプロトコルに従う情報を第 2 のプロトコルに従う情報に変換し、上記第 2 のネットワークを介して上記第 2 のアダプタに転送するストレージ装置。

【請求項 12】

第 2 のホストコンピュータと第 3 のネットワークを介して接続される第 3 のアダプタと、

上記ディスク、第 1 のアダプタ、第 2 のアダプタ、第 3 のアダプタを結合する第 2 のネットワークとを有し、

上記第1のネットワーク上の通信プロトコルが第1のプロトコルであり、上記第2及び第3のネットワーク上の通信プロトコルが第2のプロトコルである請求項11記載のストレージ装置。

【請求項13】

上記第1のプロトコルは「Ethernet II」「IEEE802.3」および「IEEE802.2」のいずれかの規格に基づくプロトコルである請求項12記載のストレージ装置。

【請求項14】

上記第1のアダプタは、

上記ホストコンピュータとファイルレベルのコマンド、データの送受信を行うファイルサーバ部と、

上記第2のアダプタとブロックレベルのコマンド、データの送受信を行うチャネルアダプタ部と、

を有し、

それらが単一のボード上または単一の筐体内に構成されている請求項12記載のストレージ装置。

【請求項15】

上記ファイルサーバ部は、

該ファイルサーバ部を制御するサーバプロセッサと、

上記第1のネットワークから来るデータ、コマンドの制御を行うLANコントローラと、

上記サーバプロセッサと上記LANコントローラを接続する第1の内部バスと、

を有し、データ、コマンドの変換処理を行い上記チャネルアダプタ部へ送り、

上記チャネルアダプタ部は、

上記変換処理を行ったデータ、コマンドを上記第2のネットワークへ送付する請求項14記載のストレージ装置。

【請求項16】

上記ファイルサーバ部と上記チャネルアダプタ部の間に、

上記データを送受信するネットワークデータバスと、

上記コマンドを送受信する制御データバスと、
を物理的または論理的に独立に有し、該ネットワークデータバスおよび制御データバスは、上記第 1 の内部バスに接続されている請求項 1 5 記載のストレージ装置。

【請求項 1 7】

上記チャネルアダプタ部は、
上記第 2 のネットワークとの上記データの送受信を制御するネットワークデータコントローラと、
上記第 2 のネットワークとの上記コマンドの送受信を制御する制御データコントローラと、
上記ネットワークデータコントローラと上記制御データコントローラを接続する第 2 の内部バスと、
を有し、該第 2 の内部バスは、上記ネットワークデータバスおよび制御データバスに接続されている請求項 1 6 記載のストレージ装置。

【請求項 1 8】

上記ファイルサーバ部は、
上記サーバプロセッサと上記第 1 の内部バスとの間の上記データ、コマンドの送受信、障害情報の制御を行うホストコントローラを有し、
上記チャネルアダプタ部は、
該チャネルアダプタ部を制御するチャネルプロセッサを有し、
上記ファイルサーバ部と上記チャネルアダプタ部の間に、上記サーバプロセッサおよびホストコントローラの少なくとも一つと、上記チャネルアダプタ部を接続し、上記第 1 の内部バスを介さずに障害情報の送受信を行う管理バスを有する請求項 1 7 記載のストレージ装置。

【請求項 1 9】

上記管理バスは上記チャネルアダプタ部内の制御データコントローラに接続されており、障害のため上記第 1 の内部バスが使えない場合でも、上記サーバプロセッサまたは上記ホストコントローラの障害情報を上記チャネルアダプタ部に転送する請求項 1 8 記載のストレージ装置。

【請求項 2 0】

上記チャネルプロセッサと接続される管理ネットワークと、
上記管理ネットワークと接続される管理プロセッサを有し、
上記チャネルプロセッサは制御データコントローラと接続されて上記障害情報を収集し、該障害情報を上記管理プロセッサに転送し、
上記管理プロセッサは受信した情報に基づいて、上記チャネルプロセッサに指示を行う請求項 1 9 記載のストレージ装置。

【請求項 2 1】

上記第 2 のネットワークに接続された制御メモリを有し、
上記管理プロセッサは、上記制御メモリにある障害情報を元に障害が発生したファイルインタフェースを持つ第 1 のアダプタの動作を停止し、他の正常なファイルインタフェースを持つ第 1 のアダプタにその制御手段の動作を代替させることを特徴とする請求項 2 1 記載のストレージ装置。

【請求項 2 2】

上記管理プロセッサは、上記チャネルプロセッサから入手した障害情報と上記制御メモリ内にある障害情報を比較し、障害が発生したファイルインタフェースを持つ制御手段の特定と、障害処理方法の選択を行うことを特徴とする請求項 2 1 記載のストレージ装置。

【請求項 2 3】

上記第 1 のアダプタは、上記ファイルサーバ部と上記チャネルアダプタ部の状態を上記管理プロセッサに送信し、上記状態を基に管理用コマンドを受信することを特徴とする請求項 2 1 記載のストレージ装置。

【請求項 2 4】

上記管理プロセッサは管理用ディスプレイと接続され、上記障害情報に基づいた情報を、管理用ディスプレイ上に表示することを特徴とする請求項 2 0 記載のストレージ装置。

【請求項 2 5】

「Ethernet II」「IEEE802.3」および「IEEE802.2」のいずれかの規格（「外部プロトコル」という）に基づくプロトコルによりホストコンピュータ

とコマンドおよびデータを送受信し、上記外部プロトコル以外のプロトコル（「内部プロトコル」という）によって記録ディスクにアクセスするディスク制御装置であって、

該制御装置は単一のボード上または筐体内にファイルサーバ部とチャネルアダプタ部を有し、

上記ファイルサーバ部は、

該ファイルサーバ部を制御し、上記外部プロトコルから内部プロトコルへデータ、コマンドの変換処理するサーバプロセッサと、

上記外部プロトコルによるデータ、コマンドの通信制御を行うLANコントローラと、

上記サーバプロセッサと上記LANコントローラを接続する第1の内部バスと、

上記サーバプロセッサと上記第1の内部バスとの間にあって、上記データ、コマンドの送受信、および、障害情報の制御を行うホストコントローラを有し、

上記チャネルアダプタ部は、

上記記録ディスクとの間で上記データの送受信を制御するネットワークデータコントローラと、

上記記録ディスクとの間で上記コマンドの送受信を制御する制御データコントローラと、

上記チャネルアダプタ部を制御するチャネルプロセッサと、

上記ネットワークデータコントローラと上記制御データコントローラを接続する第2の内部バスとを有し、

上記ファイルサーバ部と上記チャネルアダプタ部の間に、

上記第1及び第2の内部バスに接続され、上記内部プロトコルに基づいたデータを送受信するネットワークデータバスと、

上記第1及び第2の内部バスに接続され、内部プロトコルに基づいたコマンドを送受信する制御データバスとを物理的または論理的に独立に有し、

上記第1及び第2の内部バスを介さずに障害情報の送受信を行う管理バスとを有するディスク制御装置。

【発明の詳細な説明】

【 0 0 0 1 】**【発明の属する技術分野】**

本発明は、ファイルシステム及びストレージシステムを統合化した装置およびその管理方法に関し、特にファイルシステムからI/Oコマンド、障害処理を高速に処理する技術に関する。

【 0 0 0 2 】**【従来の技術】****【特許文献 1】**

特開2000-99281号公報

【特許文献 2】

特開2002-14878号公報

【非特許文献 1】

米国特許公開公報2002/0116593

従来、ホストコンピュータと接続されて用いられるストレージ装置として、特開2000-99281号公報に開示されたものがある。これは、ストレージ装置内のデータ転送、制御を高速化するために、ホストコンピュータと接続されるチャネルアダプタ、ディスク装置と接続されるディスクアダプタ、ディスクキャッシュ、制御メモリ間を相互結合網で接続する構成をとる。

【 0 0 0 3 】

また、イーサネット（登録商標）用のポートしかないユーザでも接続できるように、ストレージ装置内のファイルサーバとホストコンピュータとをイーサネットで接続し、ファイルサーバとチャネルアダプタとの間はファイバーチャネルポートのブロックインタフェースで接続したストレージ装置が、例えば米国特許出願2002/0116593に開示されている。

【 0 0 0 4 】

なお、専用のネットワークによって、例えばファイバーチャネルプロトコルによりサーバ、ストレージ間の通信を行う方式を、一般にSAN(Storage Area Network)という。また、LANに直接接続し、例えばイーサネットを介してTCP/IPプロトコルや「Ethernet II」（Ethernet は登録商標）「IEEE802.3」「IEEE8

02.2」等の規格に基づくプロトコルで通信を行う方式を一般にNAS(Network Area Storage)という。

一方、ファイルサーバから直接ストレージ装置に障害情報を伝えるパスを持つ計算機システムとして、例えば特開2002-14878号公報に開示されている方法がある。

【 0 0 0 5 】

【発明が解決しようとする課題】

本発明の課題は、SAN/NASを統合・集約化した新しいシステムを提供することにある。

【 0 0 0 6 】

図3は本発明の比較例として創作したシステム構成図である。このシステムは、相互結合網を適用し、かつ、ファイルシステム及びストレージシステムを統合・集約化したSAN/NAS統合型ストレージシステムである。

【 0 0 0 7 】

一般に、ユーザ、アプリケーションがデータのアクセスを行うためには、図3にあるようにホストコンピュータ(300)、ファイルサーバ(310)、ストレージ装置(340)を用いる必要がある。その場合ユーザ、アプリケーションはホストコンピュータ(300)に存在し、そこからネットワークやイーサネットポート(305)を経由して、ファイルを管理するファイルサーバ(310)へデータのアクセス要求を出す。ファイルサーバ(310)はユーザ、アプリケーションの把握しているデータ(ファイル)の管理を行い、ディスク(160, 161)に対してデータ(ブロック)のリード、ライト要求を出す。ファイルサーバ(310)はこのときファイル情報からブロック情報への変換を行う。ディスク(160, 161)はファイルサーバ(310)からのコマンドに従ってデータの格納、読み出しを行う。

このようなストレージ装置(340)において、ディスク(160, 161)の容量の増加、インターフェイス性能やプロセッサ性能の向上、LSIやボードの実装密度増加に伴い、ハードの集約化を行うことが進められている。それに伴いチャネルアダプタ数、ディスク数は増加し、ストレージ装置内のデータ転送、制御を

高速化するためにチャネルアダプタ（1 1 0, 3 1 0）、ディスクアダプタ（1 4 0、1 4 1）、ディスクキャッシュ（1 3 2）、制御メモリ間（1 3 1）を相互結合網（1 3 0）で接続する構成をとることが有利である。相互結合網を持つストレージ装置として、前述の特開2000-99281号公報に開示されている方法がある。

【0 0 0 8】

一方管理コストの削減の要求から、複数のユーザデータのストレージ装置（3 4 0）への集約化が進んでいる。しかしすべてのユーザがファイバーチャネルに代表されるブロックインターフェイスを持っているわけではなく、多くのユーザはイーサネットに代表されるIPインターフェイスしか持っていない場合が多い。また離れた地域よりストレージ装置（3 4 0）に接続する場合、ファイバーチャネルを用いるとコストが増大する問題がある。よって図3に示すように、ストレージ装置（3 4 0）の集約化を進めるにあたり、ストレージ装置（3 4 0）内にファイルサーバを内蔵し、イーサネットポート（3 0 5）等のIPインターフェイスしかないユーザでも接続できるようにストレージ装置内（3 4 0）にファイルサーバ（3 1 0）を置き、チャネルアダプタ（3 3 0）との間はファイバーチャネルポート（3 2 0）等のブロックインタフェースで接続したストレージ装置（3 4 0）の要望が高くなっている。

【0 0 0 9】

このようにIPインターフェイスを持つNASとして、例えば米国特許出願2002/0116593に開示されている方法がある。当該特許出願に開示されているストレージ装置は、ユーザに対してはIPインターフェイスを制御するファイルサーバを持ち、ストレージ装置とは相互スイッチを介したファイバーチャネルで接続される構成をとっている。この方法によれば、ファイルサーバ処理とストレージ処理を別々に行うので、高性能化が図れる。

【0 0 1 0】

一方、ストレージ装置の規模が増大し、NASが一般になってきた場合、システム全体の高信頼性を保つためにファイルサーバ（3 1 5）とストレージ装置（3 4 0）の間の障害処理連携が必要になってくる。特にファイルサーバ側（3 1 5

）の障害時にその情報がストレージ装置（340）側に直接伝わらなければ、障害部位の切り離し、他の部位へのフェイルオーバーなどに時間がかかることになり、信頼性、性能ともに低下する。そのために通常のデータが通るパスとは別に障害情報を別のパスを使って障害情報を送信する方式が検討されている。このような、ファイルサーバから直接ストレージ装置に障害情報を伝えるパスを持つ計算機システムとして、例えば特開2002-14878号公報に開示されている方法がある。当該公報に開示されている計算機システムは、ファイルサーバ側のプロセッサが制御するホストバスを使わず、ストレージ装置に障害情報を連絡するパスを別に持つ構成を採用している。この方法によれば、ファイルサーバ側の障害をプロセッサの状態によらず直接知ることができ、システムの高信頼化を保つことができる。

【0011】

しかし、上記した特開2000-99281号公報に開示されているストレージ装置では、ホストコンピュータとのインターフェイスはIPインターフェイスを含んでいない。従って、ファイルサーバを持っていないユーザはこのストレージ装置に接続ができない。そのため、このストレージ装置に接続するために新たにファイルサーバを用意する必要がある。図3は新たにファイルサーバ（310）を備えた例であるが、この比較例には管理コスト、設置面積が増大する問題がある。

【0012】

また米国特許出願2002/0116593に開示されているNAS装置については、装置内にNASを持つので、ファイルサーバとストレージ装置が連結されているが、1回のデータやコマンドを転送するために使う接続パスはファイバーチャネルケーブル1本であり、負荷が高い場合性能低下を招く可能性が高い。さらにファイバーチャネルケーブルのみで接続されているため、ファイルサーバに障害が発生した場合、ストレージ装置にその障害情報を伝えることができない。よってフェイルオーバーを完了するために多くの時間がかかる可能性が高い。つまり、このNAS装置では性能低下と信頼性低下の問題がある。

【0013】

一方、障害処理に関して特開2002-14878号公報に開示されている計算機システ

ムは、直接ディスク制御装置側（本公知例では I/O プロセッサと記載）にファイルサーバ（本公知例では主プロセッサと記載）の障害情報の送信ができるが、その情報を障害処理に用いるためには、ファイルサーバの設定情報を変更し IP ネットワークを通して外部の管理サーバに伝えられた後になるため、IP ネットワークの負荷が高い場合などには相手は的確に障害情報を取得できないか、障害情報の伝達に時間がかかる問題がある。

以上の課題により、本発明の目的は、ストレージ装置内にファイルサーバのインターフェイスを持つシステムおよびその方法を提供することである。

【0014】

本発明の他の目的は、ファイルサーバとストレージ装置間において、コマンドとデータを並列に処理するシステムおよびその方法を提供することである。

【0015】

本発明のさらに他の目的は、ファイルサーバの障害情報を通常のコマンド、データと別にストレージ装置に転送し、その情報をストレージ装置全体で共有するシステムおよびその方法を提供することである。

【0016】

本発明のさらに他の目的は、ストレージ装置で共有した障害情報を使って障害の発生したファイルサーバのフェイルオーバーを行うシステムおよびその方法を提供することである。

【0017】

【課題を解決するための手段】

上記課題を解決するために、本発明の一態様であるストレージ装置では、データを記憶するディスクと、第 1 のホストコンピュータに接続される第 1 のネットワークに直結される第 1 のアダプタと、ディスクに直結される第 2 のアダプタと、第 1 のアダプタ、第 2 のアダプタに直結してこれらを結合する第 2 のネットワークとを有し、第 1 のアダプタは、第 1 のネットワークを介してホストコンピュータから受信したのプロトコルに従う情報を第 2 のプロトコルに従う情報に変換し、第 2 のネットワークを介して第 2 のアダプタに転送する。本発明では、ホストコンピュータとアダプタ間を、例えば P C I インタフェースなどの内部バスを

用いて、冗長なプロトコル変換を行わない高速インタフェースで接続することができる。特に、本発明では、ファイルサーバと、ストレージ装置のチャンネルアダプタとを、第1のアダプタとして同一基板上に置き、それらの間を高速なインターフェイスで接続できるようにする。

【0018】

第1のプロトコルは、例えば「Ethernet II」「IEEE802.3」および「IEEE802.2」のいずれかの規格に基づく、いわゆるイーサネットプロトコルであり、種々のホストコンピュータとの接続を可能とする。第2のプロトコルは例えばファイバーチャンネルであり、専用の高速チャンネルを実現することができる。本発明は、第1のアダプタのみによりこの両者を結合することが可能である。従って、スペースファクターや保守に優れる。

【0019】

ファイルサーバ部は、ファイルサーバ部を制御するサーバプロセッサと、第1のネットワークから来るデータ、コマンドの制御を行うLANコントローラと、サーバプロセッサと上記LANコントローラを接続する第1の内部バスとを有する。たとえば、サーバプロセッサはデータ、コマンドの変換処理を行い、第2のプロトコルによりチャンネルアダプタ部へ送る。チャンネルアダプタ部は、変換処理後のデータ、コマンドを第2のネットワークへ送付する。

【0020】

このように、ファイルサーバ部とチャンネルアダプタ部を一体構成としたことにより、両者の間に、データを送受信するネットワークデータバスと、コマンドを送受信する制御データバスとを物理的または論理的に独立に備えることとするとも容易にできる。すなわち、ファイルサーバとストレージ装置の間に複数の独立したバスを設け、I/O処理の中でデータとコマンドを別のバスで処理できるようにする。

【0021】

さらに、ファイルサーバ部は、サーバプロセッサと第1の内部バスとの間のデータ、コマンドの送受信、障害情報の制御を行うホストコントローラを有してもよい。また、チャンネルアダプタ部は、チャンネルアダプタ部を制御するチャンネルプ

ロセッサを有してもよい。ファイルサーバ部とチャネルアダプタ部の間には、第1の内部バスを介さずに障害情報の送受信を行う管理バス備えることが望ましい。この管理バスは、例えばサーバプロセッサあるいはホストコントローラと、チャネルアダプタ部内の制御データコントローラを直結する。障害のため第1の内部バスが使えない場合でも、サーバプロセッサまたは上記ホストコントローラの障害情報を上記チャネルアダプタ部に転送することができる。このように、ファイルサーバの障害処理情報をI/O処理と異なるバスを使ってストレージ装置に転送できるようにする。また上記障害処理情報をストレージシステムで共用できるようにメモリに保存し、フェイルオーバーを行う他のストレージ装置、ファイルサーバから参照できるようにする。

【0022】

本発明の他の観点は上述の新規な第1のアダプタであり、これは、例えば「Ethernet II」「IEEE802.3」および「IEEE802.2」のいずれかの規格（「外部プロトコル」という）に基づくプロトコルによりホストコンピュータとコマンドおよびデータを送受信し、外部プロトコル以外のプロトコル（「内部プロトコル」という）によって記録ディスクにアクセスするディスク制御装置であって、制御装置は筐体内または好ましくは単一のボード上にファイルサーバ部とチャネルアダプタ部を有する。

【0023】

ファイルサーバ部は、ファイルサーバ部を制御し、外部プロトコルから内部プロトコルへデータ、コマンドの変換処理するサーバプロセッサと、外部プロトコルによるデータ、コマンドの通信制御を行うLANコントローラと、サーバプロセッサと上記LANコントローラを接続する第1の内部バスと、サーバプロセッサと第1の内部バスとの間であって、データ、コマンドの送受信、および、障害情報の制御を行うホストコントローラを有する。

【0024】

チャネルアダプタ部は、記録ディスクとの間でデータの送受信を制御するネットワークデータコントローラと、記録ディスクとの間でコマンドの送受信を制御する制御データコントローラと、チャネルアダプタ部を制御するチャネルプロセ

ッサと、ネットワークデータコントローラと制御データコントローラを接続する第2の内部バスとを有し、ファイルサーバ部と上記チャネルアダプタ部の間に、第1及び第2の内部バスに接続され、内部プロトコルに基づいてデータを送受信するネットワークデータバスと、第1及び第2の内部バスに接続され、内部プロトコルに基づいてコマンドを送受信する制御データバスとを物理的または論理的に独立に有し、第1及び第2の内部バスを介さずに障害情報の送受信を行う管理バスとを有する。

【0025】

さらに他の本発明の態様であるストレージ装置は、データを記憶する複数のディスクと、ホストコンピュータからのI/O命令を受信し、I/O命令に従って上記ディスクを制御する複数の制御装置と、複数の制御手段を結合する相互結合網とを有し、制御装置の1つは、ホストコンピュータとのインタフェースとして、ブロックインタフェースを持ち、制御装置の別の1つはホストコンピュータとのインタフェースとして、ファイルインタフェースを持つことを特徴とする。ここで、ファイルインタフェース（ファイルシステムインタフェース）とは、ファイル名を元にしてデータの送受信を行うインタフェースをいう。ブロックインタフェース（ブロックデバイスインタフェース）とは、SCSIに代表されるようなデバイス識別子、先頭ブロックアドレス、ブロック数などを元にデータの送受信を行うインタフェースをいい、これは、ディスク内のデータ位置を示すブロックアドレスを指定してデータにアクセスを行う。このように、取り扱うプロトコルの異なるインタフェースを統合したストレージ装置を提供する者である。

【0026】

【発明の実施の形態】

以下、本発明の実施例を図1を参照して説明する。

【0027】

図1は本発明の第1の実施例の構成図である。

【0028】

図1において、100はユーザ、アプリケーションのデータを保存するストレージ装置である。本発明のストレージ装置（100）は、ホストコンピュータ（

101、300)に適合するインターフェイスを持つ。ファイバーチャネルインターフェイスを持つホストコンピュータ(300)に対しては、複数のファイバーチャネルポートからなるファイバーチャネルインターフェイス(103)で接続する。IPインターフェイスを持つホストコンピュータに対しては、複数のイーサネットポートからなるイーサネットインターフェイス(104、105)で接続する。ストレージ装置(100)の内部は、ホストコンピュータ(300)のファイバーチャネルインターフェイス(103)を接続しデータ、コマンド処理を行う、チャネルアダプタ(110)と、ホストコンピュータ(101)のイーサネットインターフェイス(104)を外部ネットワーク(102)経由で接続しファイルを含むデータ、コマンドの処理を行う、ファイルサーバボード(112、115)を持つ。ファイルサーバボード(112、115)は自らOSを持ち、ホストコンピュータ(101)からのIPプロトコルの処理、ファイルレベル要求のブロックレベル要求への変換を行うファイルサーバ部(113)とブロックレベル要求を処理するチャネルアダプタ部(114)からなる。このファイルサーバ部とチャネルアダプタ部(114)は同一ボード上に配置され、後で説明するように複数の高速なインターフェイスで接続されている。またストレージ装置(100)は、データの一時的保存とデータのリード、ライトの高速化を行うディスクキャッシュ(132)とストレージ内のデータの整合性を維持、装置内の状態を保存、共用するための制御メモリ(133)を持つ。さらにストレージ装置(100)は最終的にデータを保存するディスク(160、161)を制御するディスクアダプタ(140、142)を持つ。チャネルアダプタ(110)、ファイルサーバボード(112)、ディスクキャッシュ(132)、制御メモリ(133)ディスクアダプタ(140、142)は相互結合網(130)とのインターフェイスを持ち、お互いに結合されている。相互結合網(130)はスイッチ等で構成されており、外部ネットワーク(102)に比べ高速、かつ高信頼なネットワークになっている。以上の構成により本発明のストレージ装置は、同一装置で複数のインターフェイスを持つので、図3に示したように余計なファイルサーバを用意する必要がなく、装置コストを下げる事が可能となる。次にファイルサーバボード(112、115)の詳細について説明する。

【0029】

図2は本発明におけるファイルサーバ部とチャネルアダプタ部を統合した、ファイルサーバボードの構成図である。

【0030】

図2において、200は外部ネットワーク(102)を通して、ホストコンピュータとファイルレベルのコマンド、データの送受信を行い、相互結合網(130)を通して、ディスクキャッシュ(132)、制御メモリ(133)ディスクアダプタ(140、142)とブロックレベルのコマンド、データの送受信を行う、ファイルサーバボード(112、115)の基本構成を示している。なお図2では説明を容易にするため、外部ネットワーク(102)、相互結合網(130)とのインターフェイス数は1個に限定してあるが、それらが複数存在する場合も当然本特許の範囲にある。201はファイルサーバ部であり、外部ネットワーク(102)からのコマンド、データを受け取り、自ファイルシステム内の処理を行った後、ブロックレベルのデータ、コマンドの変換処理を行いチャネルアダプタ部(202)へ送る。210はサーバプロセッサであり、ファイルサーバ部(201)の全体を制御する。211はホストコントローラであり、サーバプロセッサ(210)と周辺のメモリ、内部バス(214)のデータ、コマンド、割り込み信号の送受信、障害情報の制御を行う。212はLANコントローラであり、外部ネットワーク(102)から来るデータ、コマンドの制御を行う。213はイーサネットであり、外部ネットワーク(102)とファイルサーバボード(200)を接続し、IPプロトコルに準拠したデータ、コマンドの伝送を行う。214は内部バスであり、ホストコントローラ(211)、LANコントローラ(212)、チャネルアダプタ部(202)を接続し、データ、コマンドの送受信を行う。内部バス(214)以外に、サーバプロセッサ(210)、ホストコントローラ(211)は、障害情報の送受信を独立に行うバスとして、管理バス(230)を持つ。管理バス(230)はチャネルアダプタ部(202)内の制御データコントローラ(221)に接続されており、障害のため内部バス(214)使えない場合でも、サーバプロセッサ(210)、ホストコントローラ(211)の障害情報をチャネルアダプタ部(202)に転送する。202はチャネル

アダプタ部であり、ファイルサーバ部（201）から送られた、ブロックレベルのデータ、コマンドを受け取り、内容に応じて相互結合網（130）を通して適当なディスクキャッシュ（132）、制御メモリ（133）、ディスクアダプタ（140、142）に送信する。220はチャンネルプロセッサであり、上記チャンネルアダプタ部（202）の全体を制御する。221は制御データコントローラであり、ストレージ装置（100）全体を制御するために必要なデータの送受信の制御、サーバプロセッサ（210）、ホストコントローラ（211）からの障害情報の送受信の制御を行う。222は制御メモリ（133）用の相互結合網（130）であり、制御データコントローラからの制御データを制御メモリ（133）、ディスクアダプタ（140、142）、他のチャンネルアダプタ（110）、ファイルサーバボード（113）へ転送する。223はネットワークデータコントローラであり、ユーザまたはアプリケーションのデータの送受信を制御する。224はユーザ、アプリケーションデータ転送用の相互結合網（130）であり、ネットワークデータコントローラ（223）からのデータをディスクキャッシュ（132）、ディスクアダプタ（140、142）に転送する。231は制御データバスであり、コマンドやストレージ装置（100）全体を制御するために必要な情報をファイルサーバ部（201）とチャンネルアダプタ部（202）間で送受信する。232はネットワークデータバスであり、ユーザまたはアプリケーションのデータやデータ転送に必要なパラメータ等をファイルサーバ部（201）とチャンネルアダプタ部（202）間で送受信する。250は管理ネットワークであり、チャンネルプロセッサ（220）で収集した障害情報、構成情報を管理プロセッサ（255）に転送するために利用する。管理プロセッサ（255）は受信した情報より、障害処理、構成変更等を判断し、必要であればストレージ装置（100）に指示を行う。また管理プロセッサ（255）は各ボードの情報を管理ディスプレイ（260）に一元的に表示することが可能である。たとえば261のような管理表を表示することが可能であり、ボード#（265）毎にファイルサーバ部（266）、チャンネルアダプタ部（267）の状態を表示することが可能となる。240はファイルサーバボード（200）の電源でありここから、ファイルサーバ部（201）、チャンネルアダプタ部（202）に対して電源を

供給する。以上に述べた構成によって、同一ボード上にファイルサーバとストレージ装置のインターフェイスをまとめることが可能となり、装置の設置面積が削減する。またファイルサーバとストレージ装置を同一管理プロセッサ（255）、管理ディスプレイ（260）上で管理することも可能となり、管理コストが削減する。さらにファイルサーバ部（201）とチャネルアダプタ部（202）間のデータ、コマンド転送に関して、制御データパス（231）とネットワークデータパス（232）の2種類の独立したパスを設けることにより、コマンド、パラメータなどの比較的短いデータと、ユーザ、アプリケーションデータなどの比較的長いデータを分離して処理することにより、データ転送性能を上げることが可能になる。またコマンド等をまとめて転送することもでき、処理効率を上げることが可能となる。さらにコマンドとデータを分離することにより、コマンドのフォーマットを標準仕様から変更できるので装置特有な情報、パラメータをチャネルアダプタに直接伝えることも可能となり、ディスクキャッシュ（132）内の常駐データの選別などが可能となる。この2種類の独立したパスは、物理的に分離した2本のパスで構成する場合と、以下に述べるファイルサーバボード（112）のハード構成図のように物理的には1本のパスにて、論理的に2本のパスがあるように制御を行う方式のどちらを使っても実装可能であるがそれらが本特許の範囲にあることは明らかである。また図2ではファイルサーバボード（112）上にサーバプロセッサ（210）、チャネルプロセッサ（220）の2種類のプロセッサが稼動している例を挙げたが、この形態以外にサーバプロセッサ（210）のみでファイルサーバボード（112）上の部品をすべて制御する方式も実装可能であり、それも本特許の範囲であることは明らかである。

【0031】

図4は本発明の実施例1におけるファイルサーバボードのハード構成図である。

【0032】

図4においてファイルサーバ部（400）内のサーバプロセッサ（210）とホストコントローラ（211）の間は、ホストバス（413）で接続されている。ホストコントローラ（211）には管理バス（555）を制御する管理バスコ

ントローラ（５１２）とサーバプロセッサ（２１０）のプログラム、データ等を格納するローカルメモリ（４１０）と内部バス（２１４、２２５）を接続し制御する内部バスコントローラ（５１１）が接続されている。内部バスコントローラ（４１１）はファイルサーバ部（４００）とチャネルアダプタ部（４６０）間を接続するバスを制御し、論理的に複数のバスがあるように内部バス（２１４、２２５）を制御する。つまり図３における制御データバス（２３１）とネットワークデータバス（２３２）を論理的に構成することが可能となる。チャネルアダプタ部（４６０）内のチャネルプロセッサ（２２０）はローカルバスコントローラ（４７０）を介してローカルバス（４７１）に接続されている。図４では省略されているが、ローカルバスにはチャネルプロセッサ（２２０）に必要なメモリ等が接続されている。またローカルバスコントローラ（４７０）は管理ネットワーク（２５０）と接続されており、必要な管理情報等を管理プロセッサ（２２５）に送信している。ローカルバス（４７１）には制御データコントローラ（２２１）が接続されている。制御データコントローラ（２２１）には内部バス（２５５）が接続されており、通常の制御データの送受信が行われる。また相互結合網（２２２）を通して制御メモリ（１３１）、他のチャネルプロセッサと接続されており、制御情報、共有情報を送受信している。さらにサーバプロセッサ（２１０）、管理バスコントローラ（２２５）からの割り込み信号（４５６、４５７）が直接入力される。また、内部バス（２１４、２２５）が障害等で稼動していない場合にもサーバプロセッサの情報が入手できるように、管理バス（４５５）も接続されている。制御データコントローラ（２２１）内部には、通常制御を行うための通常レジスタ（４８０）、エラー情報を記憶するためのエラーレジスタ（４８１）、ファイルサーバ部（４００）との通信制御を行うためのドアベルレジスタ（４８２）、通信メモリ（４８３）がある。内部バス（２２５）には制御データコントローラ（２２１）以外にネットワークデータコントローラ（２２３）が接続されており、ユーザ、アプリケーションデータを送受信する。また、チャネルプロセッサ（２２０）がファイルサーバ部（４００）の起動、停止、再起動を制御できるように、制御データコントローラ（２２１）と管理バスコントローラ（２２５）は、電源制御手段（４５８）で接続されている。本実施例では、電源



制御手段（４５８）によってファイルサーバ部の電源を制御できるものとする。
図４の構成はあくまでもひとつの実現例であり、他の構成でも本特許の範囲にあることは自明である。次に以上のべたファイルサーバボード（２００）内にある制御データバス（２３１）を使って送受信されるコマンドの構造例を示す。

【００３３】

図５は本発明の実施例１のコマンドデータブロックのフォーマットを示す図である。

【００３４】

図５において、５００はファイルサーバ部（４００）からチャネルアダプタ部（２０２）に送信されるコマンドデータブロックを示している。５０１はコマンドの種別であり、通常のコマンドか特殊なコマンドなのかを区別する。５０２、５１０、５３５、５４０はリザーブ領域で特に意味はない。５０３はTAG種別であり、キューに格納された場合の制御方法を指示する。５０４はサーバプロセッサ#であり、このコマンドを発行したサーバプロセッサ（２１０）を特定する。５０５はIIDであり、ユーザ領域またはシステム領域へのアクセスを区別する。５０６はLUNであり、サーバプロセッサ（２１０）がアクセスする先の論理ユニットの番号である。５１１はTAG#をであり、キューの内部での識別番号を示す。５１５はCDBであり、通常のSCSIで使われるフォーマットに準拠したコマンドが入る。５１６はCDB内のオペコードであり、リード、ライト等のコマンドを区別する。５２０はイニシエータポート番号であり、外部にコマンドを送信する場合の送信元のポートの区別を行う。５２５は外部デバイスWWNであり、外部にコマンドを送信する場合の送信先WWNを示す。５３０は外部デバイスLUNであり、外部にコマンドを送信する場合の送信先LUNを示す。５４１はアドレスエントリ数を示し、以下に続くストレージ装置（１００）へ指示するパラメータの個数を示す。５４５、５６０はLBA１、LBA２であり、LU内での相対的なアドレスを示す。５５０、５６５はサーバプロセッサ（２１０）のローカルメモリ（４１０）内での物理アドレスを示し、対象とするデータの格納位置となる。５５５、５７０はストレージ装置指示パラメータであり、ファイルサーバ部内に特有な情報をここに置きチャネルアダプタに伝える。５５１、５７１は各パラメータに対応するデ

ータのバッファサイズまたは転送データ長を示す。ストレージ装置指示パラメータの例としては、ディスクキャッシュ（１３２）へのデータ常駐指示がある。これを用いることにより、ストレージ装置（１００）だけでは知ることのできないユーザ、アプリケーションデータの使用頻度等をつかむことが可能となり、ストレージ装置（１００）の性能向上につなげることが可能になる。以上述べた、コマンドデータブロックが、サーバプロセッサ（２１０）、制御データコントローラ（２２１）、チャンネルプロセッサ（２２０）、ディスクキャッシュ（１３２）間でどのように送受信されるかについて以下に説明する。

【００３５】

図６は、本発明の実施例１のデータアクセス時のフローチャートを示す図である。

【００３６】

サーバプロセッサ（２１０）、チャンネルプロセッサ（２２０）はコマンドを効率よく処理するためにそれぞれコマンドキュー（CmdQueue）（６０２、６０７）、キューの先頭を示すポインタ（CmdHead）（６０１、６０６）、次に来るコマンドを格納すべき位置を示すポインタ（CmdNext）（６０３、６０８）を持つ。コマンドキュー（６０２、６０７）はリングバッファになっている。またコマンド終了後のステータスの管理のために、ステータスキュー（StQueue）（６３２、６３７）、キューの先頭を示すポインタ（StHead）（６３１、６３６）、次に来るコマンドを格納すべき位置を示すポインタ（StNext）（６３３、６３８）を持つ。ステータスキュー（StQueue）（６３２、６３７）もリングバッファとなっている。コマンド、データ転送時に、サーバプロセッサ（２１０）、チャンネルプロセッサ（２２０）間でコマンドキュー（CmdQueue）（６０２、６０７）の内容を随時確認している。両者において、CmdHead（６０１、６０６）と、CmdNext（６０３、６０８）の内容は一致している必要がある。そのため、制御データコントローラ（２２１）を通してCmdHead（６０６）、CmdQueue（６０２）、CmdNext（６０３）をそれぞれ、サーバプロセッサ（２１０）、チャンネルプロセッサ（２２０）、チャンネルプロセッサ（２２０）に転送し、整合性をとる。同様な処理はステータスキュー（StQueue）（６３２、６３７）にも必要であり、サーバプロセッ



サ (210)、チャンネルプロセッサ (220) 間でステータスキュー (StQueue) (632、637) の内容の確認を随時行う。コマンド、データ処理を行うために、サーバプロセッサ (210) のプロセス (600) で、以下、3ステップの処理が行われる。

(1) コマンドキュー (CmdQueue) (602) に空きがあることを確認後、コマンドを格納。

(2) キューポインタを更新し、CmdNext (603) を見てリングバッファが溢れないように制御。

(3) ポーリングを開始し、ステータスキューの更新をチェック。

このプロセスに対して、チャンネルプロセッサ (220) のプロセス (605) で以下の5ステップの処理が行われる。

(4) ポーリングにより、コマンドキュー (CmdQueue) (607) の更新をチェック。

(5) コマンドキュー (CmdQueue) (607) の内容をDMA転送にて、受信。

(6) ポーリングにより、ステータスキュー (StQueue) (637) の更新をチェック。

(7) キューの先頭を示すポインタ (CmdHead) (606) を更新。

(8) 各コマンドを処理。

【0037】

次にサーバプロセッサ (210) のプロセス (610) にてコマンド処理に必要なパラメータ (611) をDMA転送にて、制御データコントローラ (221) を通してチャンネルプロセッサ (220) に転送する。また同時にサーバプロセッサ (210) のプロセス (620) にてデータをDMA転送にて、ネットワークデータコントローラ (223) を通してディスクキャッシュ (132) に転送する。プロセス (610) とプロセス (620) は、図3、4で示したように独立したパスにて処理できるので、並列に処理を進めることが可能である。データ転送処理が完了した後、チャンネルプロセッサ (220) のプロセス (635) にて以下の2ステップの処理が行われる。

(9) 実行結果をステータスキュー (StQueue) (637) に格納。



(10) ドアベルレジスタ (582) セットし割り込みを発生。

このプロセスに対して、サーバプロセッサ (210) のプロセス (630) で、以下の 3 ステップの処理が行われる。

(11) 割り込みまたはポーリングより、ステータスキュー (StQueue) (632) の更新を検知

(12) ステータスの取り出し

(13) キューのポインタ (StNext) (633) の更新

以上のフローチャートに示したように、コマンド、ステータス毎にキューを持つことが可能となり、また各々、独立のバスを用いて処理できるので、データアクセス性能を向上することが可能となる。

【0038】

図 7 は、本発明の実施例 1 のサーバプロセッサ障害発生時のフローチャートを示す図である。

【0039】

図 7 において、700 はサーバプロセッサ (210) に障害が発生し、例外処理を行う必要があることを示す。ここで、通常コマンド、データが送受信される、内部バス (214、225) と内部バスコントローラ (411) は正常に動作することが期待できない。障害情報はプロセス (810) にて例外ハンドラの処理が進み、エラー報告がプロセス (715) にて制御データコントローラ (221) のエラーレジスタ (482) に割り込み信号で記録される。その後サーバプロセッサ (210) はリセット命令を待つことになる。次に制御データコントローラ (221) はプロセス (715) でサーバプロセッサ (210) に障害が発生していることをエラー割り込みの形でチャネルプロセッサ (220) に伝える。チャネルプロセッサ (220) のプロセス (720) ではエラー割り込みを受けてエラー処理関数が起動する。その後チャネルプロセッサ (220) はサーバプロセッサ (210) を初期化するために、プロセス (715) にて制御データコントローラ (221) の通常レジスタ (483) にリセット命令を送信する。制御データコントローラ (221) は次にサーバプロセッサ (210) にリセットを割り込み信号にて送信する。リセットを受けたサーバプロセッサ (210)

のプロセス（710）は次のプロセス（725）にて再起動を行う。プロセス（725）ではサーバプロセッサ（210）内の障害要因を収集して、プロセス（730）にて制御データコントローラ（221）の通信メモリ（481）に格納する。ただしこの障害要因は通常バスを通して転送されるので、正しい値である保障はない。プロセス（725）はこれらの情報の格納が終了した後、退避完了報告をドアベルレジスタ（482）経由でチャネルプロセッサ（220）に伝える。その後、チャネルプロセッサ（220）は正しい障害情報を確保するために、プロセス（750）、プロセス（715）において管理バス（455）経由で障害情報を読み出す。具体的にはチップセットレジスタリードを行う。また同時にプロセス（745）にて制御データコントローラ（221）の通信メモリ（481）に格納された障害要因も読み出す。これらの情報を読み出した後、チャネルプロセッサ（220）は管理プロセッサ（255）へ情報を報告し、処理指令を受領する。次に制御メモリ（131）へ障害情報転送し、他のボードに障害処理要求を出す。次にプロセス（755）にて内部バス（214、225）のエラーが発生しているかどうかをチェックする。エラーがなければ、チャネルプロセッサ（220）はプロセス（770）にてメモリダンプの指示をネットワークデータコントローラ（223）に対して行う。ネットワークデータコントローラ（223）は、ローカルメモリ（410）の情報等をプロセス（765）にてDMA転送にてメモリダンプする。もし内部バス（214、225）にエラーが発生していた場合、メモリダンプは行わない。次にチャネルプロセッサ（220）は外部からの障害発生ファイルサーバボード停止要求に応じて、プロセス（790）にて制御データコントローラ（221）に対して強制停止命令を発行する。制御データコントローラ（221）は電源制御手段（458）によって、サーバプロセッサ（210）を強制停止させる機能を持っており、上記強制停止命令を受けて、サーバプロセッサ（210）は強制閉塞する。以上の処理では、通常のバス以外に障害情報を送信するバスを設けているので、バスに障害があった場合でも障害情報をチャネルプロセッサ（220）に対して送信でき、障害処理の高信頼化を図ることが可能となる。図7では、単一ファイルサーバボード（112）内の障害情報の通信方法を示したので、これを用いて複数ファイルサーバボード（

112) 間でのフェイルオーバーを行う方法について以下で説明する。

【0040】

図8は、本発明の実施例1のサーバプロセッサ障害時のフェイルオーバーのフローチャートを示した図である。800は図7同様、サーバプロセッサ(210)に障害が発生し、例外処理を行う必要があることを示す。障害を感知したチャネルプロセッサ(220)は図7で説明したサーバプロセッサ障害発生時の処理を行う。その中で制御メモリ(131)に対して障害情報と障害処理要求を書き込む。その後フェイルオーバー先のファイルサーバボード(115)がプロセス(810)にて障害処理を検知する。その後ファイルサーバボード(115)は障害発生元のファイルサーバボード(112)に対して停止要求を出す。その後停止完了の通知が来るまでプロセス(815)の待ち状態に入る。ファイルサーバボード(115)からのファイルサーバボード停止要求を受けた、ファイルサーバボード(112)は図7で説明したように、サーバプロセッサを停止する。さらに、プロセス(820)で残っているI/O処理を破棄した後、制御メモリ(131)に対してファイルサーバボード(112)の停止を書き込み、チャネルプロセッサを停止する。ファイルサーバボード(112)の停止を受けた制御メモリ(131)はプロセス(825)にてストレージ装置(100)の構成情報を更新し、停止をフェイルオーバー先のファイルサーバボード(115)に通知する。停止の通知を受けた、フェイルオーバー先のファイルサーバボード(115)はプロセス(930)にてファイルサーバの引継ぎを行い、その結果を管理プロセッサ(255)に通知し、フェイルオーバーを完了する。

【0041】

以上述べた実施例により本特許に開示されたSAN/NAS統合型ストレージ装置を使うと、通常のI/O処理では、サーバプロセッサ(210)、チャネルプロセッサ(220)毎にコマンドキュー(CmdQueue)(602、607)、ステータスキュー(StQueue)(632、637)を持つことにより、非同期にコマンド処理、ステータス処理を進めることが可能になり、またコマンド、データを独立のパスを用いて処理することも可能になるので、I/O処理性能を向上できる。さらに、コマンド、データで独立なパスを使い、またコマンドを拡張することにより

、ファイルサーバ固有のI/O特性をストレージ装置に伝えることが可能となり、ディスクキャッシュ（132）のヒット率の向上等のストレージ装置（100）最適化が図れる。さらに同一ボード上にファイルサーバ、ディスク制御装置のチャンネルアダプタがあり、同一の管理プロセッサ（255）、管理ディスプレイ（260）で管理することが可能になるので、設置面積の減少、管理コストの削減ができる。障害処理では、管理バス（455）から障害情報を送信し、ストレージ装置（100）で共有している制御メモリ（131）に書き込むことで、バスの障害にかかわらず、他のファイルサーバボードに対して早く障害情報を伝えることが可能となり、フェイルオーバーの効率を上げ、信頼性を向上できる。

【0042】

（実施例2）

図9は、本発明の実施例2のサーバプロセッサ障害時のフェイルオーバーのフローチャートを示す図である。

図8では、障害発生元のファイルサーバボード（112）が自立的に障害発生をフェイルオーバー先のファイルサーバボード（115）に通知したが、障害のよって通知ができない場合があり、障害処理を開始できない。このようなケースを避けるため、フェイルオーバー先のファイルサーバボード（115）が積極的に他のファイルサーバボード（112）を監視する方式が考えられる。正常な場合、ファイルサーバボード（112）は一定時間間隔でハートビートをフェイルオーバー先のファイルサーバボード（115）に伝える。もしプロセス（900）のように、障害が発生して一定時間内にこのハートビートが通知されない場合、フェイルオーバー先のファイルサーバボード（115）は障害が発生と判断し、プロセス（905）のように障害処理を開始する。障害発生元のファイルサーバボード（112）が障害情報を制御メモリ（131）に書き込んだ後、フェイルオーバー先のファイルサーバボード（115）は制御メモリ（131）に構成情報を読み出し、障害が発生したことを確認し障害処理内容を確定する。その後プロセス（905）にてファイルサーバボード（115）は障害発生元のファイルサーバボード（112）に対して停止要求を出す。その後の処理は図8と同様に、その後停止完了の通知が来るまでプロセス（920）の待ち状態に入る。フ

ファイルサーバボード（１１５）からのファイルサーバボード停止要求を受けた、ファイルサーバボード（１１２）は図７で説明したように、サーバプロセッサを停止する。さらに、プロセス（９２５）で残っているI/O処理を破棄した後、制御メモリ（１３１）に対してファイルサーバボード（１１２）の停止を書き込み、チャンネルプロセッサを停止する。ファイルサーバボード（１１２）の停止を受けた制御メモリ（１３１）はプロセス（９３０）にてストレージ装置（１００）の構成情報を更新し、停止をフェイルオーバー先のファイルサーバボード（１１５）に通知する。停止の通知を受けた、フェイルオーバー先のファイルサーバボード（１１５）はプロセス（９３５）にてファイルサーバの引継ぎを行い、その結果を管理プロセッサ（２５５）に通知し、フェイルオーバーを完了する。

以上述べた実施例により、開示されたSAN/NAS統合型ストレージ装置を使うと、障害発生元のファイルサーバボード（１１２）から障害発生の送信ができない場合でも早い段階でフェイルオーバー先のファイルサーバボード（１１５）が障害発生の検知、確認を行うことが可能となり、フェイルオーバーの効率を上げ、信頼性を向上できる。

【００４３】

【発明の効果】

本発明によれば、ファイルサーバとディスク制御装置間のI/Oコマンド、データの送受信を独立して行えるため、I/O処理性能が向上できるという効果がある。またアプリケーション、ファイルシステムのI/O特性情報をディスク制御装置に与えることができるため、I/O処理の最適化ができるという効果がある。また同一のボード上にファイルサーバとディスク制御装置のチャンネルアダプタを実装し、統一した管理プロセッサ、管理プロセッサで管理するために、装置の設置面積の減少、管理コストの削減ができる効果がある。また、独立した経路でファイルサーバの障害情報を共有できるため、フェイルオーバーの効率化、高信頼化ができるという効果がある。

【図面の簡単な説明】

【図１】

本発明のSAN/NAS統合型ストレージシステムの構成図。

【図 2】

本発明におけるファイルサーバ部とチャネルアダプタ部を統合した、ファイルサーバボードの構成図。

【図 3】

比較例のSAN/NAS統合型ストレージシステムの構成図。

【図 4】

実施例 1 におけるファイルサーバボードのハード構成図。

【図 5】

実施例 1 のコマンドデータブロックのフォーマットを示す図。

【図 6】

実施例 1 のデータアクセス時のフローチャートを示す図。

【図 7】

実施例 1 のサーバプロセッサ障害発生時のフローチャートを示す図。

【図 8】

実施例 1 のサーバプロセッサ障害時のフェイルオーバーのフローチャートを示す図。

【図 9】

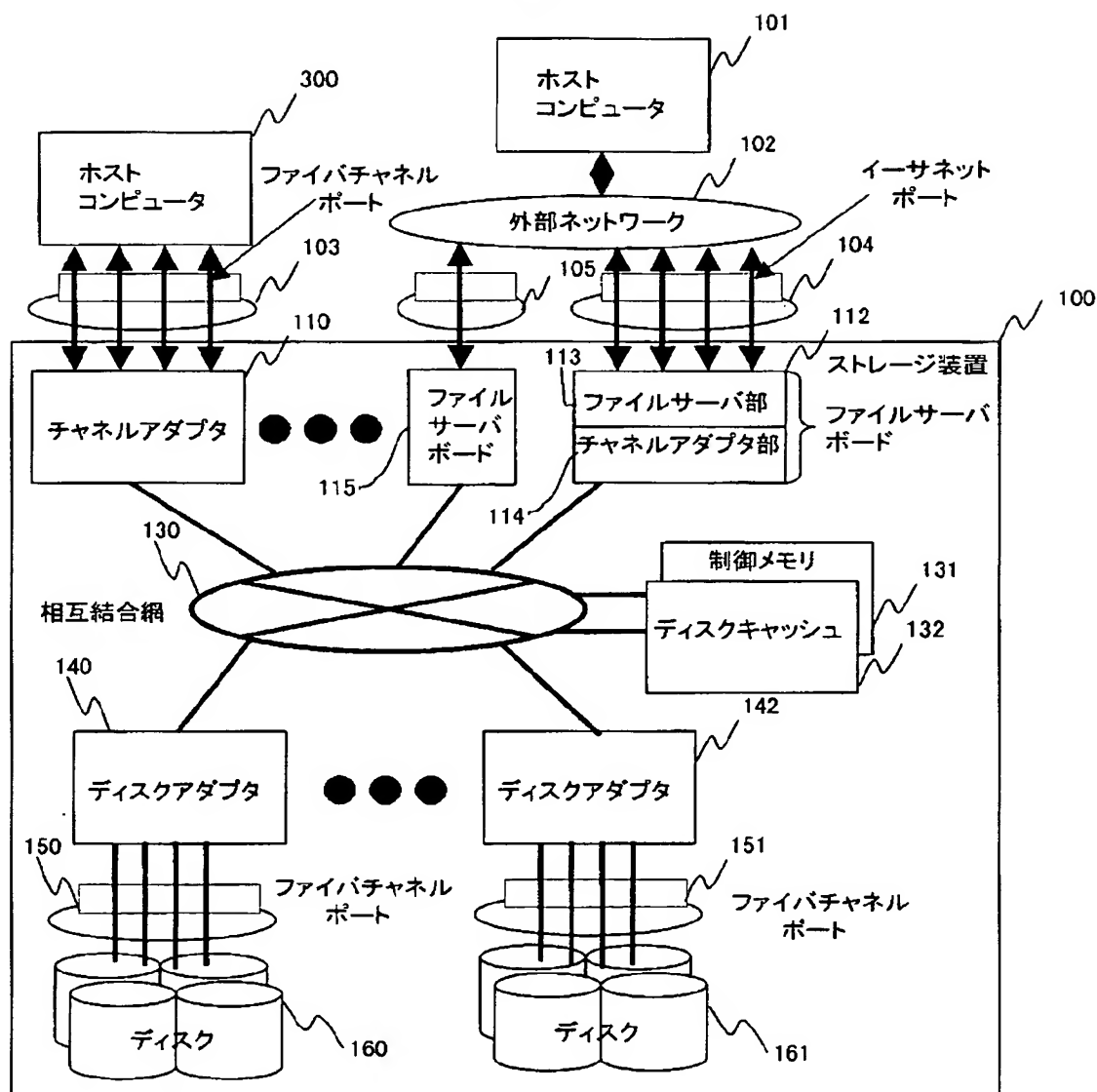
実施例 2 のサーバプロセッサ障害時のフェイルオーバーのフローチャートを示す図。

【符号の説明】

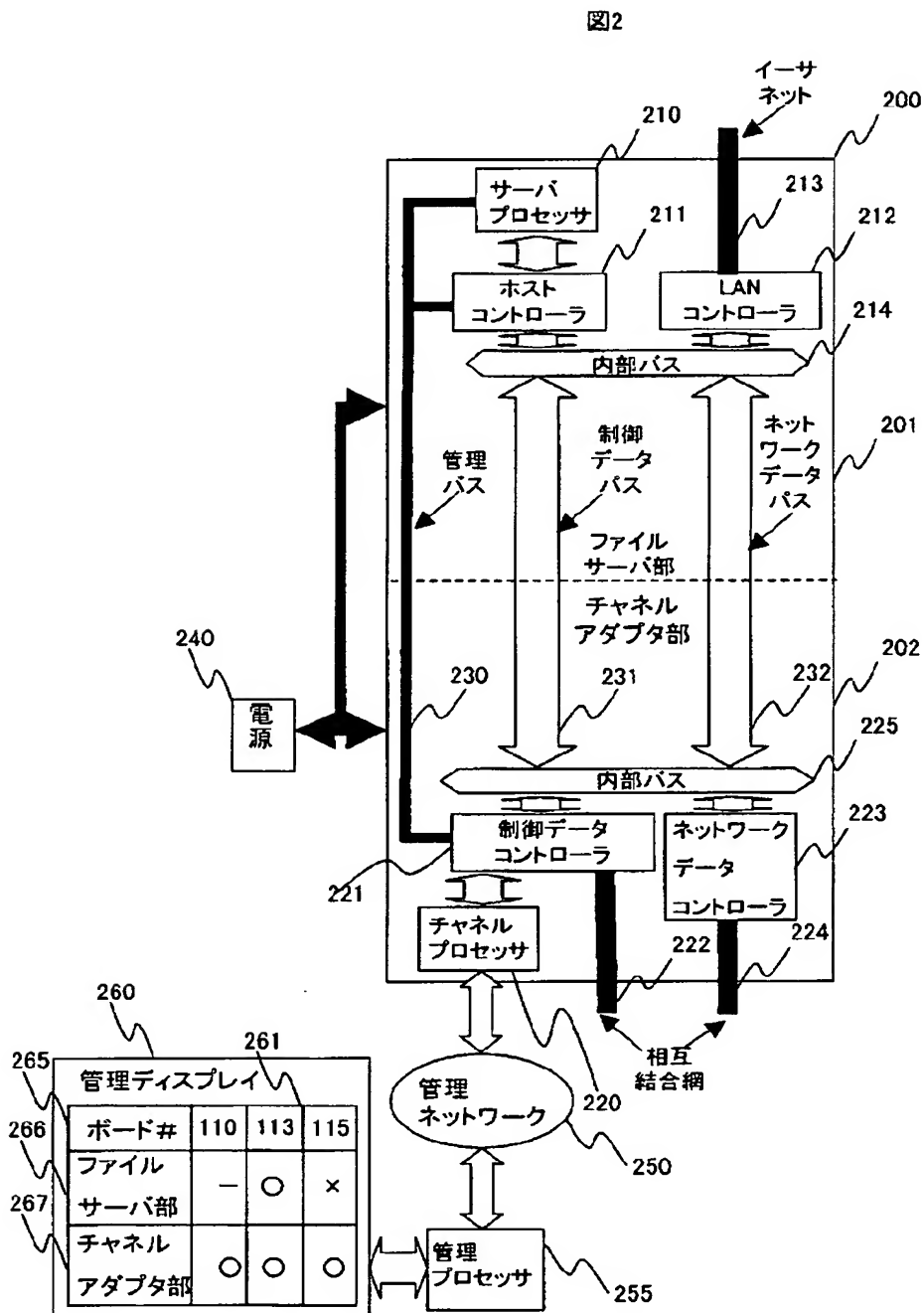
100、101、150、151：ファイバチャネルインターフェイス、 102：イーサネットインターフェイス、 110：チャネルアダプタ、112：ファイルサーバボード、113：ファイルサーバ部、114：チャネルアダプタ部、 130：相互結合網、 131：ディスクキャッシュ、 132：制御メモリ、 140、142：ディスクアダプタ、 160、161：ドライブ。

図面

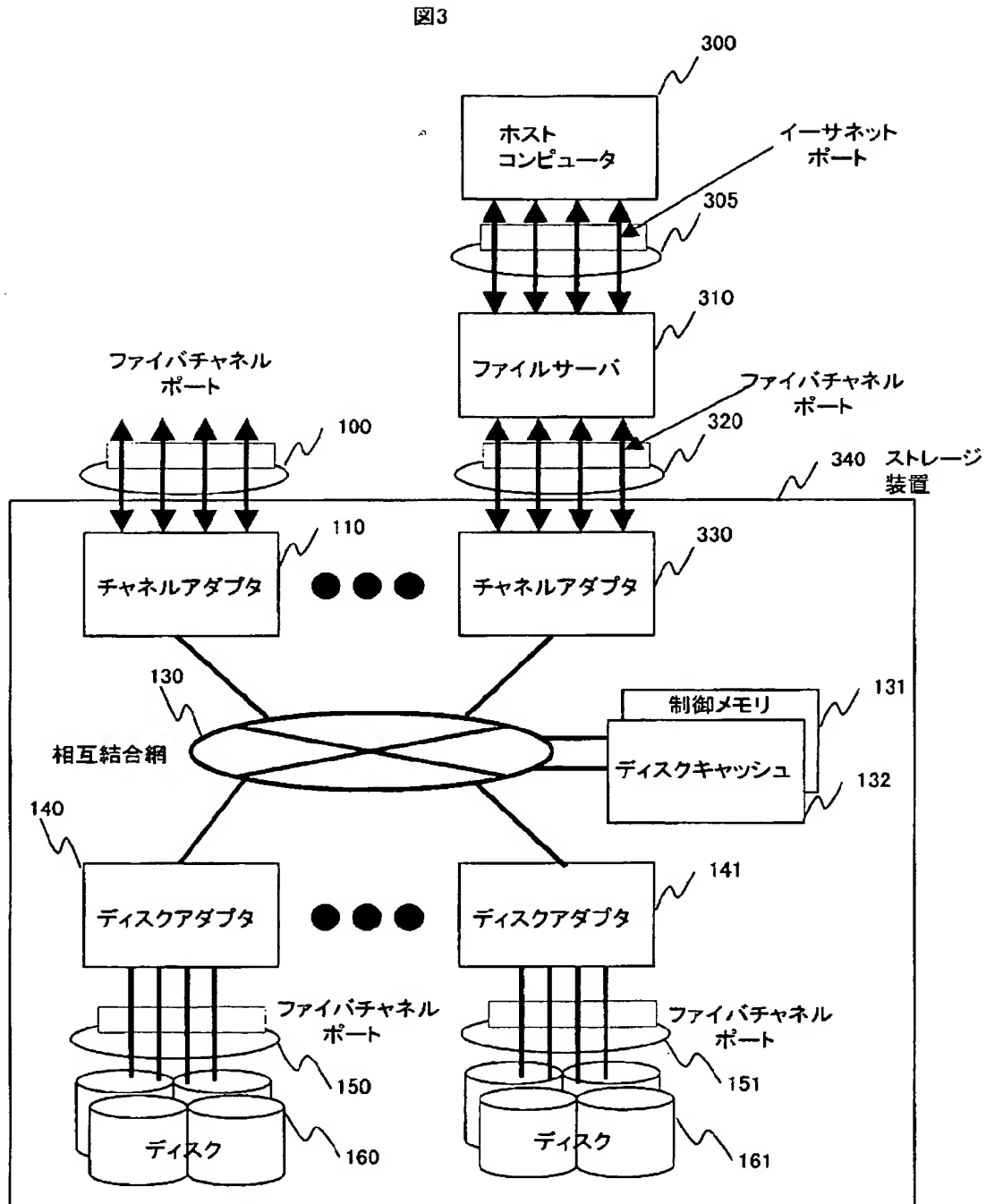
图 1



【図 2】

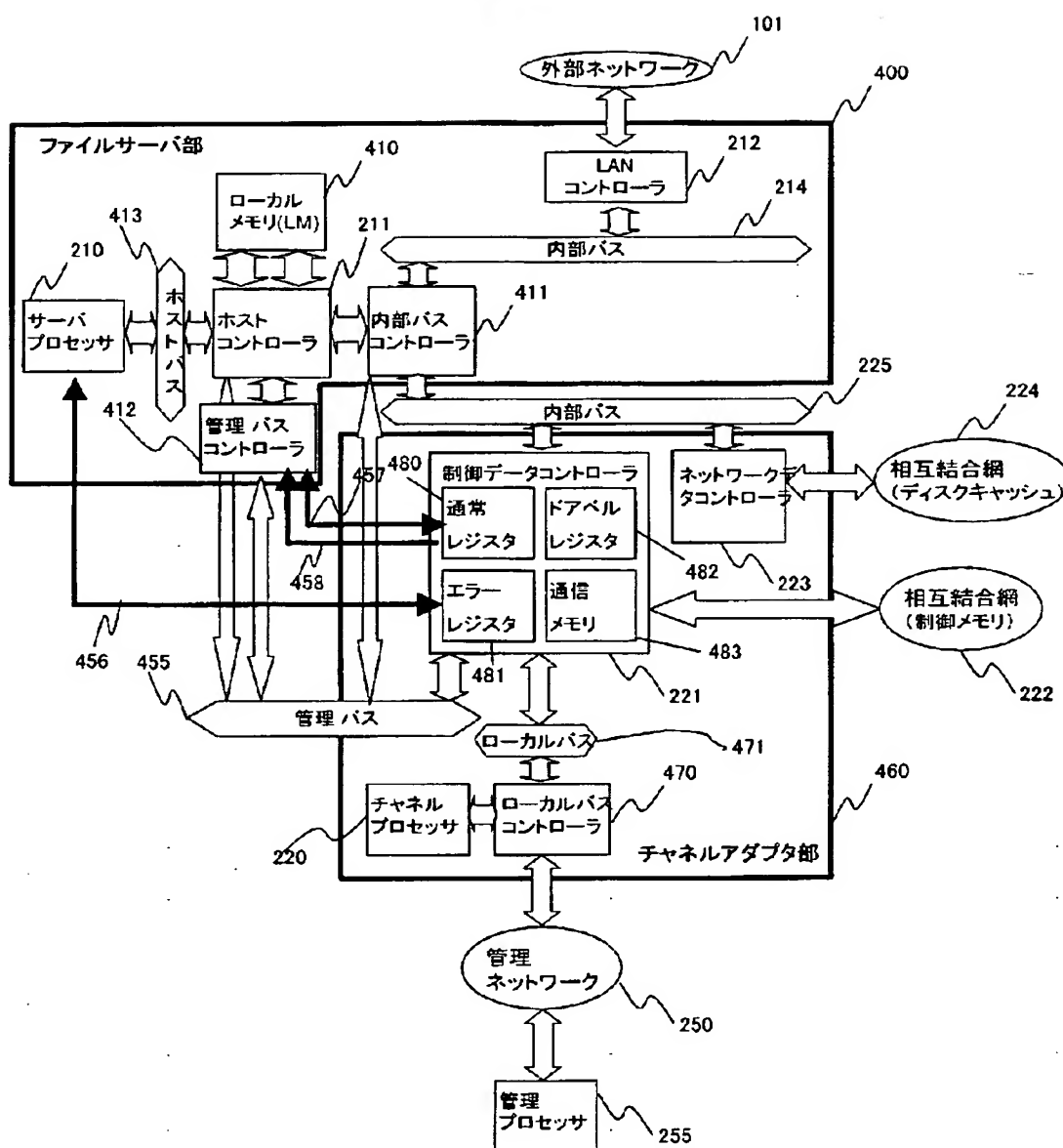


【図 3】



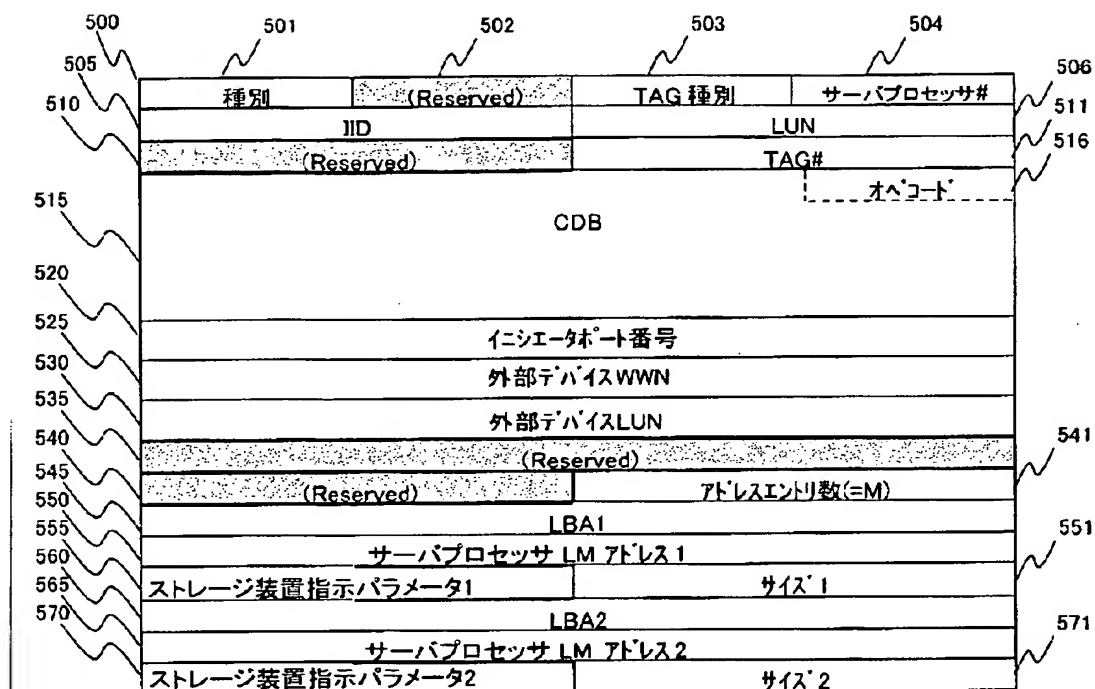
【図 4】

図4

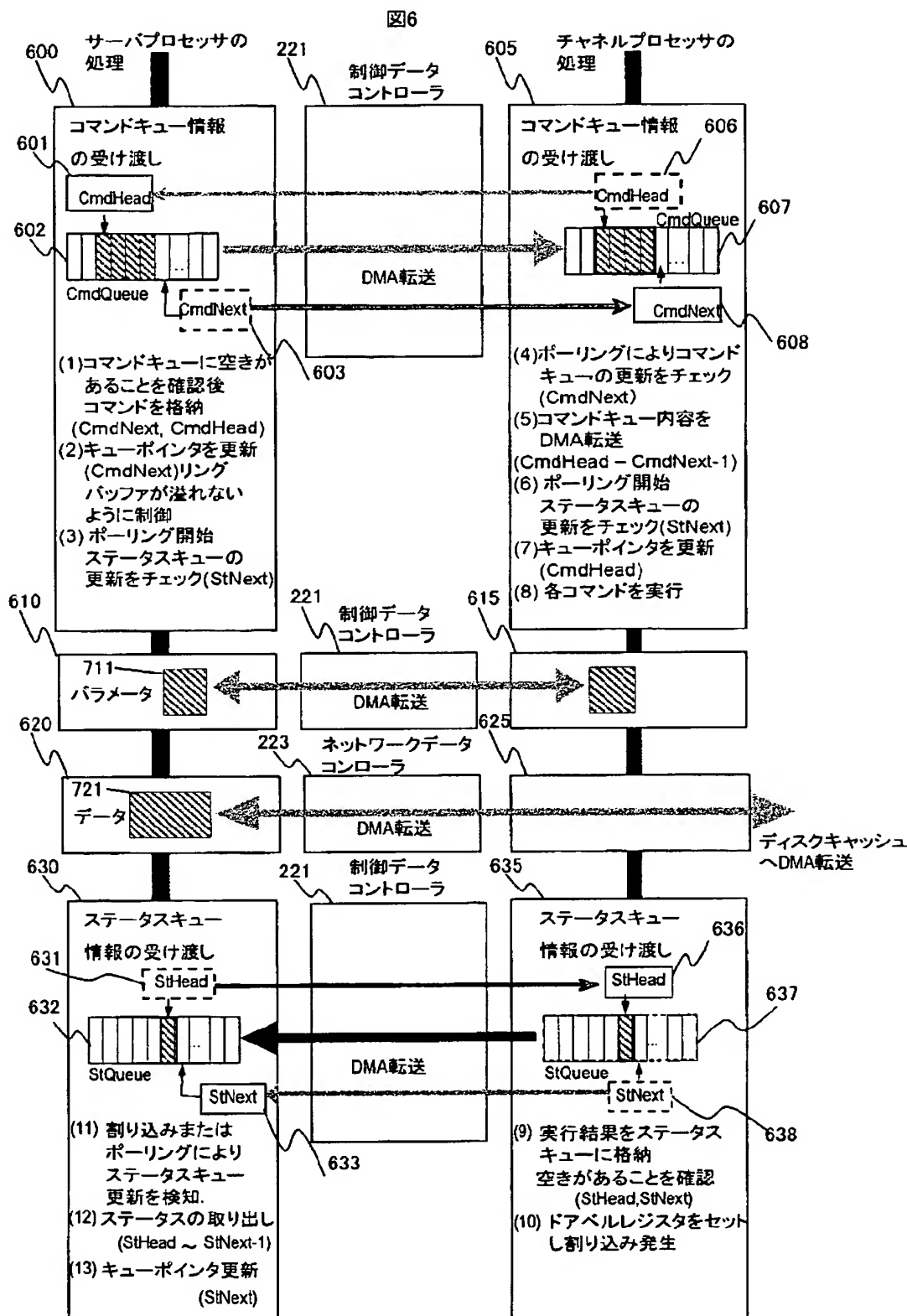


【図 5】

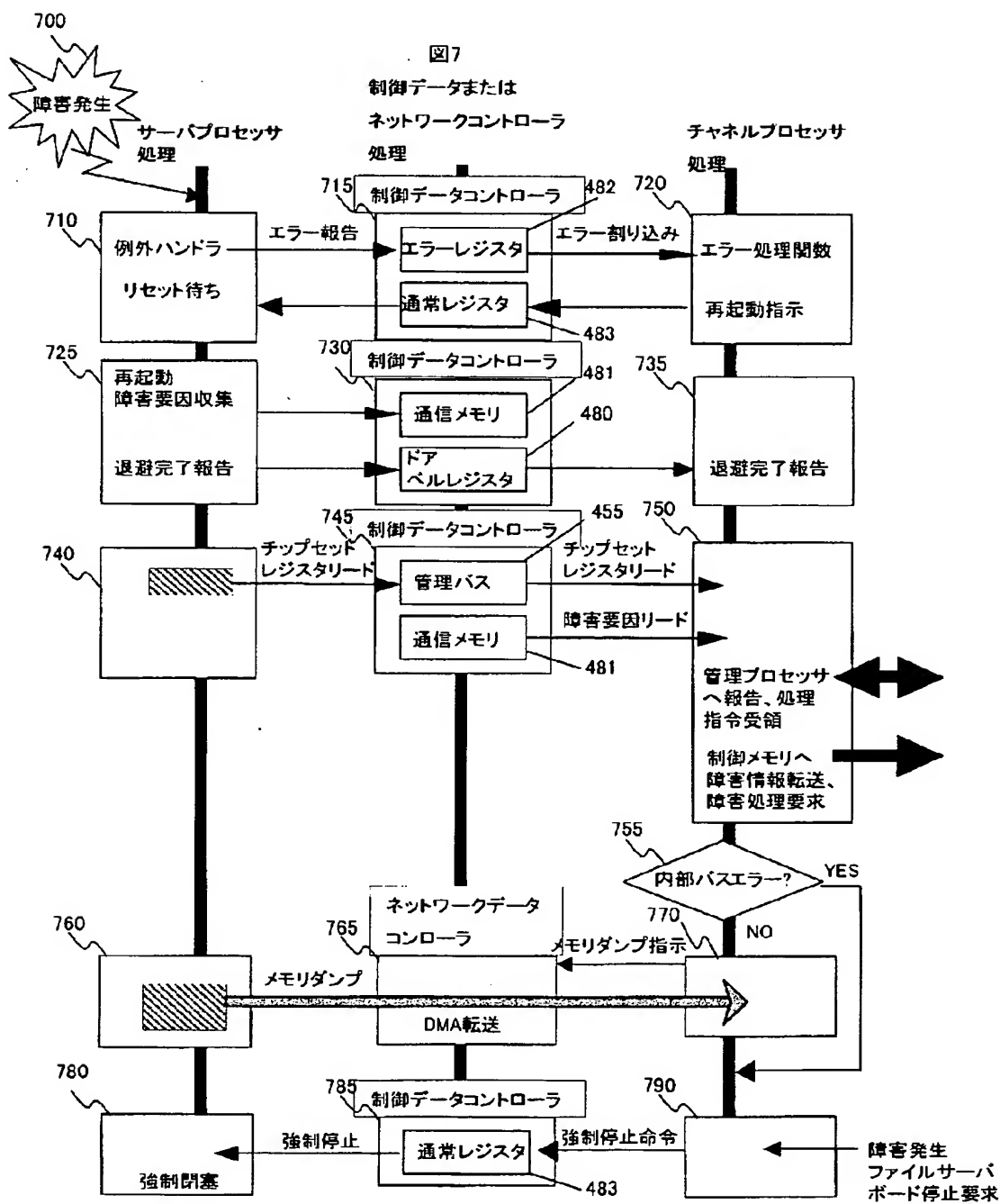
図5



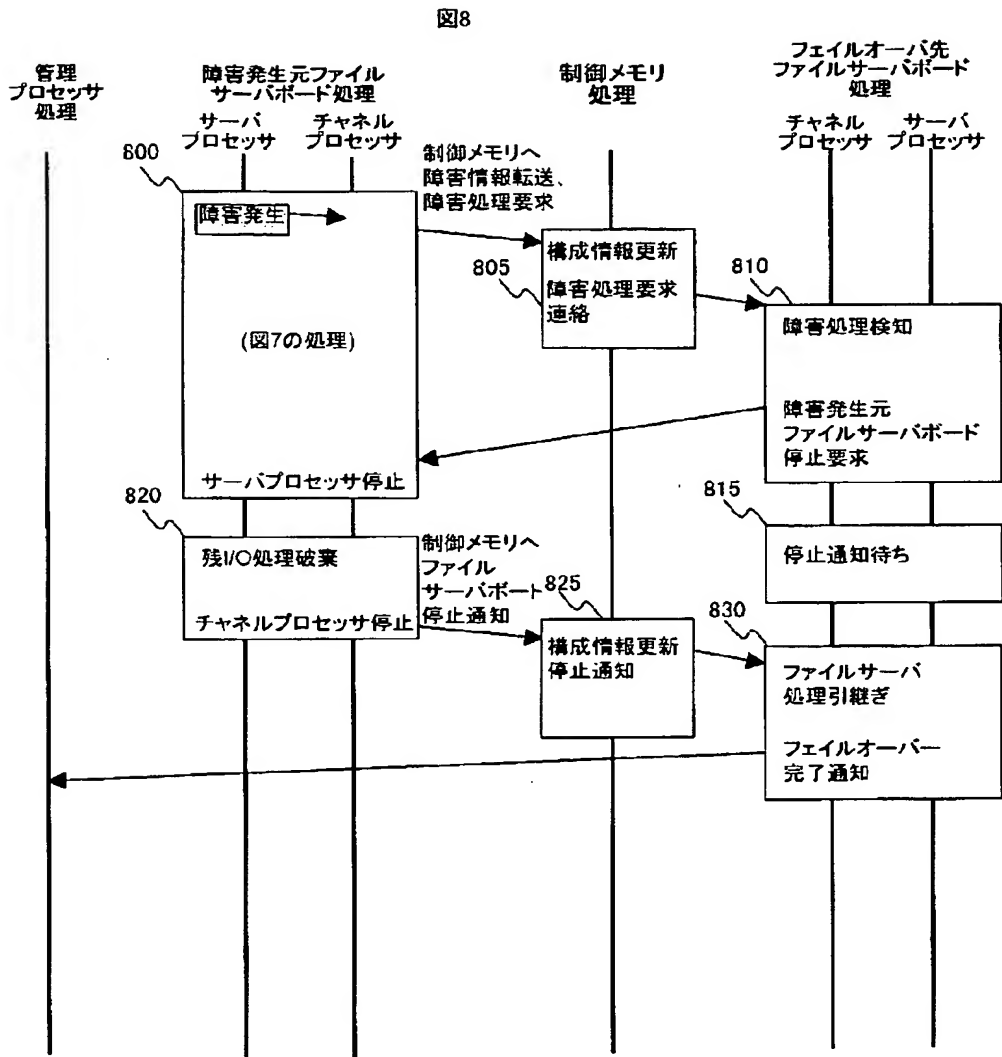
【図 6】



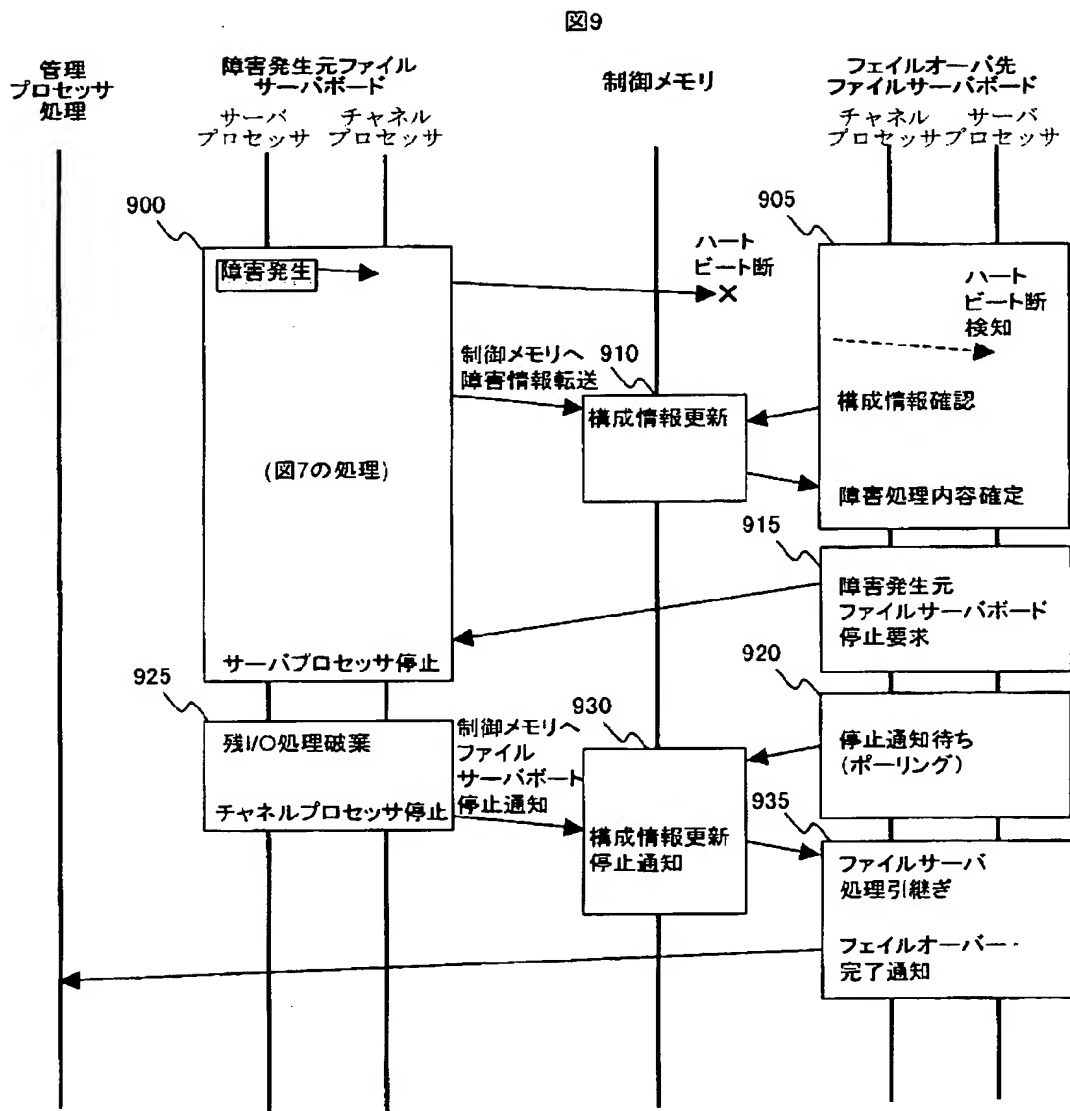
【図 7】



【図 8】



【図 9】



【書類名】 要約書

【要約】

【課題】 ネットワークに直結するストレージシステムにおいて、ファイルサーバで発行された、I/Oコマンドをストレージ装置に送信する際に、従来のI/F及び、プロトコルを用いると、コマンド、データを単一の転送路にてシリアルに転送するために、性能低下が生じる。またアプリケーション、ファイルシステムに固有のI/O特性をストレージ装置に伝達することができない。一方、ファイルサーバのフェイルオーバーを高速かつ正確に行うために、ファイルサーバ側の障害をストレージ装置側で共有する必要もあるが、従来のI/Fではそれを伝える手段がない。

【解決手段】 ファイルサーバとストレージ装置内のチャネルアダプタを同一ボード112上に配置し、その間の接続パスが、複数のプロトコルを独立して稼動できるように制御を行う。またファイルサーバ113とチャネルアダプタ114の間でコマンド、データのパスと独立したパスを置き、受信した障害情報をストレージ装置100内の共有メモリに格納し、フェイルオーバーに用いる。

【選択図】 図1

認定・付加情報

特許出願の番号	特願 2003-005234
受付番号	50300037782
書類名	特許願
担当官	第七担当上席 0096
作成日	平成15年 1月15日

<認定情報・付加情報>

【提出日】 平成15年 1月14日

次頁無

特願 2003-005234

出願人履歴情報

識別番号

[000005108]

1. 変更年月日

1990年 8月31日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台4丁目6番地

氏 名

株式会社日立製作所